CN 53-1189/P ISSN 1672-7673

基于 MPI 的高性能 UVFITS 数据合成研究与应用*

陈泰燃[1]，王 威[3]，王 锋[1,2]，邓 辉[1]，刘应波[1]，梅 盈[1]

（1. 昆明理工大学云南省计算机技术应用重点实验室，云南 昆明 650500；2. 中国科学院云南天文台，云南 昆明 650012；3. 中国科学院国家天文台，北京 100012）

摘要：中国明安图超宽频谱射电日像仪（Mingantu Ultrawide Spectral Radioheliograph，MUSER）进入实际观测后，每 3 ms 产生一帧 100 kB 左右的数据，每天的原始观测数据约 3.5 TB。由于射电日像仪的原始数据采用自定义格式，为了后续数据分析和共享的需要，有必要根据数据存储需求把这些原始数据转换成天文常用的文件格式。在前期工作中已经实现了原始数据格式到 UVFITS 文件的转换，在此基础上研究了基于 MPI 的集群并行环境下 UVFITS 合成系统的性能优化。通过实验验证，在改进后的并行环境下，UVFITS 合成系统的性能达到了需求的 2.5 倍，可以有效处理当前及未来一定时间内射电日像仪的海量观测数据。同时，改进后的系统具有良好的横向扩展能力，能够为相关项目的数据处理提供借鉴和参考。

关键词：UVFITS；海量数据；MPI；并行计算；高性能计算

中图分类号：TP311.1 **文献标识码**：A **文章编号**：1672-7673(2016)02-0184-06

中国明安图超宽频谱射电日像仪（Mingantu Ultrawide Spectral Radioheliograph，MUSER）是同时以高时间、空间和频率分辨率对太阳进行射电频谱成像的设备[1]。随着进入系统联调和试观测阶段，当前迫切需要实现试观测结果的存储、处理，以推进设备误差修正、设备定标等一系列工作。这其中，需要应用当前射电天文领域的相关软件，如天文应用软件公共包（Common Astronomy Software Applications，CASA）[2]等对试观测数据进行处理与分析，从而获得同行公认的结果。如何将观测得到的结果快速保存为天文应用软件公共包可以接受的数据格式成为当前一个急需解决的问题。

在射电天文领域，数据存储广泛采用随机组（Random groups）结构的 FITS 文件[3]，因此主要保存 UV 复可见条纹数据也经常被称为 UVFITS 文件。

目前，MUSER-I 以 3 帧/秒的速率产生数据，每产生 8 帧数据后利用 1 ms 的时间用于数据传送。故每 25 ms 产生 8 帧的数据，如此循环，直到 1 min 生成一个大的数据文件。该文件包含观测的详细数据信息，大小将近 2 GB，其中包含 19 200（8 × 40 × 60）帧。由于其中间隔 8 帧产生的数据出自同一天线，故可根据需求，对 8 帧的数据进行积分操作以提高数据的精确度，同时可以节约存储空间。速度上，需要达到每 3 ms 处理一帧数据生成 UVFITS[4] 文件，才能满足实时生成观测图像的要求。

目前国外类似的项目有日本野边山宇宙观测所（The Nobeyama Radio Observatory，NRO）[5]，法国南锡射电望远镜（The Nançay Decimetric Radio Telescope，NRT）等，MUSER-I 的数据规模远高于国外现有项目。MUSER-I 共有 40 面天线，时间分辨率为 3 ms，信道为 16 个。在日本野边山宇宙观测所，天线数量为 84 面，但时间分辨率低。相比之下 MUSER-I 的数据量大约是日本野边山宇宙观测所的 20 倍，是南锡射电望远镜的 37 倍。同时在图像处理方面也有不同的需求，这导致国外的日像仪处理软

* 基金项目：国家自然科学天文联合基金（U1231205）；国家自然科学基金（11263004，11203011，11163004，11103005）；云南省应用基础基金重点项目（2013FA013，2013FA032，2013FZ018）资助.

收稿日期：2015-06-17；修订日期：2015-07-07

作者简介：陈泰燃，男，硕士. 研究方向：分布式计算，科学数据处理. Email：584680399@qq.com

通讯作者：王 锋，男，教授. 研究方向：分布式数据存储与计算，科学数据处理. Email：wf@cnlab.net

件无法直接引入。所以，为了更好地支撑射电日像仪的后续处理工作，设计并实现一套能够为射电日像仪服务的高性能 UVFITS 合成系统是当前比较迫切的工作。

本文研究了 UVFITS 的合成机制，对串行 UVFITS 合成系统中能够并行化的环节进行并行化处理，在当前广泛使用的成熟的消息传递接口(Message Passing Interface，MPI)基础上，结合具体应用设计并实现了一套基于消息传递接口的并行数据合成系统，用于满足射电日像仪的需求。

1　UVFITS 合成流程

日像仪产生的原始数据文件包括以下信息：每一帧的时间、信号中心频率选择开关、信号中心频率、信号带宽、量化电平值与阈值、延迟调整开关、条纹旋转开关以及子带工作方式，同时还记录许多时延信息、天线参数等。为了满足对数据存储的需要，在 UVFITS 文件中需要 4 个二元表：PRIMARY、ANTENNA、FREQUENCY、SOURCE。针对产生的数据文件，为了提高数据的精确度，需要对数据进行积分操作，系统会按照指定的起始时间和截止时间要求对数据进行积分操作。

为了查阅、检索数据的方便，把处理后的数据保存为 UVFITS 格式的文件。整个过程及 UVFITS 文件的合成处理流程如图 1。

天线接收的射电噪声信号经过低噪声放大器放大后，再进行电/光转换，将电信号转换为光信号并通过光纤传至机房[6]。经过模拟接收机和数字接收机处理后，最终生成约定格式的数据并以文件的方式存储[7]。

下一步，UVFITS 合成系统将存储的原始数据读入内存，然后读取其中的时间、日期等数据，计算当前天线所在位置。把这些处理结果存入 UVFITS 文件的 AIPS AN 表。合成系统计算观测时太阳的位置，存入 AIPS SU 表中。合成系统读取天线的极化、接收频率等信息存入 AIPS FR 表。最后，处理原始数据中的可视数据(图像)，存入 PRIMARY 表。

可视图像数据来源于原始数据中的可视数据。综合孔径成像原理要求天线接收的两路信号同时同相到达相关器[8]，在实现过程中，需要对系统延时和相位进行补偿，这就是延时补偿和条纹停止技术[9]。在生成 UVFITS 文件中 PRIMARY 表时，根据输入的时间以及其他参数，系统需要对可视数据进行处理，如果是多帧合成，需要积分，如不合成，转存即可。在这步操作中，对每一帧的时间、位置以及可视数据等信息进行处理。最后，将 4 个表(AIPS AN、AIPS SU、AIPS FR、PRIMARY)中的数据，添加上若干 UVFITS 文件的关键字存入文件。

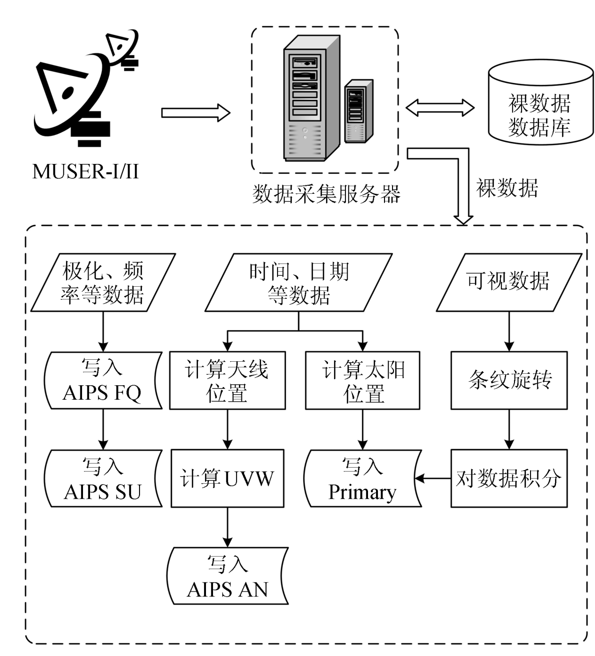


图 1　UVFITS 合成流程

Fig. 1　The UVFITS assembly flow

2　UVFITS 合成处理并行化分析

由于单机性能有限，为了尽可能大地提升程序性能，并具有进一步提升的空间，选择多机并行的方式，并且使用其中具有代表性的消息传递接口作为多机并行的方案。

通过分析，程序的整个运行过程，大致可以归纳为输入运行参数、读取数据、处理数据、输出数据。其中处理数据又可以分为分析数据和整理、积分数据。

在程序中可以做到并行的部分有读取数据、处理数据和输出数据。在测试中发现，程序的时间开销主要集中在读取数据以及整理、积分数据两个过程。其中程序读取数据这部分占用了整个程序运行时间的 83%左右，处理数据中的整理、积分数据大约占了整个程序运行时间的 12%。

在初步验证性并行测试中发现，对数据的整理、积分进行并行处理的效果并不理想。在多机并行条件下，数据的通信时间接近数据处理的时间，导致增加节点对性能的提升几近于无。而单机并行的情况下，由于总线速度的瓶颈，单核与多核的运行时间相差无几。

据此，在程序的读取数据，即读取原始数据中的帧部分，采用并行处理。

2.1 数据的并行化处理

2.1.1 方案设计与数据划分

本系统使用的方案，在集群环境中采用主从模式实现可视数据的并行处理。主要执行顺序为，主节点将等待读取的原始数据以帧为处理单元，划分为若干任务，按顺序将任务分发给每个子节点；每个子节点处理完接收的任务后将处理完毕的文件写给主节点。

在方案设计中，对帧的划分以一帧为一个单位，按照节点数的不同，平均分配到每个节点上由各节点读取。在每个参与计算的子节点上，各存有一份完整的帧头信息。在各子节点内，再将数据按初始设定的进程数进一步划分，并在进程内处理完毕。各子节点产生的结果文件直接写入主节点的存储介质，各子节点间无需通信。并行部分的流程示意图如图 2。

2.1.2 结果合并

各节点接收的任务均为至少一个完整的帧，同时每帧存有各自的时间信息。完成任务后，计算结果由子节点直接写入主节点的存储介质中保存为文件，文件名为各节点数据中读取的时间。由于任务间相互独立，合并结果的开销仅与主节点存储介质的输入/输出性能有关。

图 2 并行部分程序执行流程图

Fig. 2 The parallel execution flow of the UVFITS assembly system

3 性能分析

并行 UVFITS 文件的合成过程相对于顺序的 UVFITS 文件合成步骤，加入了数据在节点间的分发与收集。这样与串行的过程相比，额外增加了节点间通信的开销。

本节的性能分析以原始数据读入内存后的时间为基准，性能分析的标准为加速比以及并行环境下程序所耗费的时间。下面为两种方式的通用表示。

3.1 相对加速比

同一并行算法，在单节点上的运行时间与在多节点构成的处理机系统上的运行时间相比，定义为相对加速比(以下简称加速比)，用来衡量并行系统或程序并行化的性能和效果。计算方法如(1)式：

$$Sp = T_{\mathrm{serial}}/T_{\mathrm{parallel}}, \tag{1}$$

其中，T_{serial}是程序在串行方式下运行耗费的时间，串行方式采用了与并行相同的算法；T_{parallel}是程序在并行环境下运行耗费的时间。

3.2 并行方式下程序运行的时间开销

这里的时间取程序读入内存后一直到执行完毕耗费的时间。由于采用了并行方式，程序额外产生了节点间通信的时间开销。同时，程序的总时间由耗费时间最长的节点决定。时间的计算如(2)式：

$$T_{parallel} = T_{bcast} + T_{send} + T_{calculate} + T_{reduce}, \quad (2)$$

其中，T_{bcast}为根节点广播数据耗费的时间，广播的数据为每个节点均需要的数据，即目前程序运行的参数。通过此参数，各节点能计算本节点所需执行的任务在裸数据中的起止位置；T_{send}为根节点点对点通信耗费的时间；$T_{calculate}$为根节点计算、处理数据耗费的时间；T_{reduce}为根节点接收时间耗费最长的节点的数据耗费的时间。

4　实验分析

测试的硬件环境为 4 台曙光天阔 A620r-G，中央处理器为 AMD Opteron 6128，主频 2.0 GHz，内存 4 GB；硬盘为希捷 ST3250310AS，7 200 转，缓存 8 MB，容量为 250 GB；操作系统为 Linux，机群数据交换的速度为 1 Gbps，采用的消息传递接口实现为 MPICH-2。在对系统的测试中，从文件名为 MUSER_20140122-131903_342668765 的原始数据中随机取 4 段观测数据，数据量分别为 415 帧、735 帧、1 055 帧、1 375 帧。4.1、4.2、4.3 为在对数据不积分和进行 5 帧积分情况下进行测试得到的结果。

4.1　单机串行

表 1 为单机串行环境下，程序在处理不同的数据量时耗费的时间。从表 1 可以发现，时间随着帧数的增加而线性增加，但时间效率远低于射电日像仪对 UVFITS 的数据存储需求的每帧 3 ms 处理时间，仅达到了目标效率的一半。

表 1　单机串行执行时间

Table 1　Stand-alone serial execution time

指标 / 帧数	总时间 /s	平均每帧 /s	5 帧积分 /s	平均每帧 /s
415	4.06	0.009 78	2.59	0.006 24
735	7.17	0.009 75	4.59	0.006 24
1 055	10.28	0.009 75	6.60	0.006 25
1 375	13.41	0.009 75	8.59	0.006 24

4.2　单机并行

表 2 是在单台服务器上，分别使用二进程、四进程以及八进程运行程序得到的时间耗费以及加速比。可以看到，节点数每增加一倍，执行时间缩短一半以上，程序并行的加速比稳定在 1 以上，即每有 N 个进程，程序的执行时间小于等于原时间的 $1/N$。同时，随着数据量的增加，即帧数的增加，程序耗费的时间呈线性上升。这说明程序的效率是可以预计的，在需要程序达到更高的速度时，可以对需要增加的进程有一个初步估计。

4.3　多机并行

表 3 是程序在 4 台服务器组成的四节点集群并行环境下，分别使用二进程、四进程、八进程运行程序得出的时间和加速比。与单机环境下的并行相比，执行时间略有增加。这是由于进程在不同机器节点间通信的时间远大于进程通过系统内部总线传递数据的通信时间，导致额外的时间开销。加速比约等于 1，这意味着每增加一倍节点，时间的开销仍然缩短接近一半。同时，总体效果达到了期望值，超出了 3 ms 处理一帧数据的需求。

5　结束语

本文针对射电日像仪对高性能 UVFITS 数据合成要求，研究了基于消息传递接口的高性能 UVFITS 合成系统。系统可以根据实际需要将观测得到的原始数据高速转换成可以导入天文应用软件公共包的标准 UVFITS 文件。目前，优化后的并行 UVFITS 合成系统在八进程的运行环境下，只需要 1.2 ms 即可处理一帧文件，已经满足当前射电日像仪对 3 ms 每帧的速率需求，性能比需求提高了 1.5 倍。另外，基于消息传递接口的并行合成系统，从实验中可以看到具有很好的扩展性，这为下一步射电日像仪更高性能的数据合成提供了可能性，同时，也能够满足未来射电日像仪产生的更多数据的实时合成处理需要。

表 2 单机并行执行时间

Table 2 Stand-alone parallel execution time

		指标 / 帧数	总时间 /s	平均每帧 /s	加速比
二进程	不积分	415	1.95	0.004 69	2.082
		735	3.45	0.004 69	2.078
		1 055	4.98	0.004 72	2.066
		1 375	6.44	0.004 68	2.082
	积分	415	1.27	0.003 06	2.039
		735	2.25	0.003 06	2.040
		1 055	3.23	0.003 06	2.043
		1 375	4.20	0.003 05	2.045
四进程	不积分	415	0.97	0.002 33	4.185
		735	1.69	0.002 29	4.242
		1 055	2.44	0.002 31	4.217
		1 375	3.16	0.002 98	4.243
	积分	415	0.61	0.001 47	4.245
		735	1.09	0.001 48	4.211
		1 055	1.58	0.001 49	4.177
		1 375	2.05	0.001 49	4.190
八进程	不积分	415	0.49	0.001 18	8.285
		735	0.84	0.001 14	8.535
		1 055	1.20	0.001 13	8.575
		1 375	1.61	0.001 17	8.329
	积分	415	0.32	0.000 77	8.093
		735	0.55	0.000 74	8.345
		1 055	0.78	0.000 73	8.461
		1 375	1.04	0.000 75	8.259

表 3 多机并行执行时间

Table 3 Multi-machine parallel execution time

		指标 / 帧数	总时间 /s	平均每帧 /s	加速比
二进程	不积分	415	2.05	0.004 94	1.980
		735	3.61	0.004 91	1.986
		1 055	5.17	0.004 90	1.990
		1 375	6.71	0.004 88	1.998
	积分	415	1.30	0.003 13	1.992
		735	2.30	0.003 12	1.995
		1 055	3.30	0.003 12	2.000
		1 375	4.30	0.003 12	1.997
四进程	不积分	415	1.02	0.002 45	3.980
		735	1.81	0.002 46	3.961
		1 055	2.58	0.002 44	3.988
		1 375	3.37	0.002 45	3.979
	积分	415	0.64	0.001 54	4.046
		735	1.14	0.001 55	4.026
		1 055	1.64	0.001 55	4.024
		1 375	2.14	0.001 55	4.014
八进程	不积分	415	0.49	0.001 18	8.285
		735	0.86	0.001 17	8.337
		1 055	1.24	0.001 17	8.298
		1 375	1.66	0.001 20	8.078
	积分	415	0.33	0.000 79	7.848
		735	0.56	0.000 76	8.196
		1 055	0.83	0.000 78	7.951
		1 375	1.07	0.000 77	8.028

在未来的工作中进一步研究如下问题：(1)继续优化数据处理流程，提高并行度，从而提升合成 UVFITS 文件的速度；(2)在并行环境中，提前对多机进行文件同步以达到流水线效果，从而减少子节点对主节点存储介质访问的通信时间；(3)为数据处理以及归档查询系统提供应用程序编程接口 (Application Programming Interface, API)，满足数据访问的需要。

参考文献：

[1] Yan Y, Zhang J, Wang W, et al. The Chinese Spectral Radioheliograph—CSRH [J]. Earth, Moon, and Planets, 2009, 104(1): 97-100.

[2] Jaeger S. The Common Astronomy Software Application (CASA) [C] // Astronomical Data Analysis Software and Systems XVII. 2008: 623.

[3] Wells D C, Greisen E W, Harten R H. FITS-a flexible image transport system [J]. Astronomy and Astrophysics Supplement Series, 1981, 44: 363.

[4] 高姣姣，王锋，戴伟，等. 面向射电日像仪的随机组结构剖析与文件设计 [J]. 天文研究与技术——国家天文台台刊，2013，10(4)：365-371.

Gao Jiaojiao, Wang Feng, Ji Kaifan, et al. An analysis of the random-group data format and a design of the data file structure for a solar radio heliograph [J]. Astronomical Research & Technology

——Publications of National Astronomical Observatories of China, 2013, 10(4): 365-371.

[5] Nakajima H, Nishio M A, Enome S, et al. The Nobeyama radioheliograph [J]. Proceedings of the IEEE, 1994, 82(5): 705-713.

[6] 王威，陈志军，姬国枢，等. CSRH 光纤传输方案探讨 [J]. 天文研究与技术——国家天文台台刊, 2006, 3(2): 143-147.
Wang Wei, Chen Zhijun, Ji Kaifan, et al. Optical fiber transmission analysis for CSRH [J]. Astronomical Research & Technology——Publications of National Astronomical Observatories of China, 2006, 3(2): 143-147.

[7] 姬国枢，窦玉江，王威，等. CSRH 模拟接收机设计 [J]. 天文研究与技术——国家天文台台刊, 2006, 3(2): 135-142.
Ji Guoshu, Dou Yujiang, Wang Wei, et al. RF receiver design for CSRH [J]. Astronomical Research & Technology——Publications of National Astronomical Observatories of China, 2006, 3(2): 135-142.

[8] 张坚，颜毅华，刘飞，等. 用于双天线干涉实验的数字相关接收机 [J]. 天文研究与技术——国家天文台台刊, 2006, 3(2): 148-153.
Zhang Jian, Yan Yihua, Liu Fei, et al. A prototype digital correlation receiver for two-element interferometer experiment [J]. Astronomical Research & Technology——Publications of National Astronomical Observatories of China, 2006, 3(2): 148-153.

[9] 刘东浩，颜毅华，赵岸，等. 新一代厘米—分米波射电日像仪系统延时校准方法研究 [J]. 电子学报, 2013, 31(3): 570-574.
Liu Donghao, Yan Yihua, Zhao An, et al. A delay calibration for Chinesespectral radioheliograph in the decametric to centimetric wave range [J]. Acta Electronica Sinica, 2013, 31(3): 570-574.

The Study and Application of a High Performance UVFITS Assembly System Based on MPI

Chen Tairan[1], Wang Wei[3], Wang Feng[1,2], Deng Hui[1], Liu Yingbo[1], Mei Ying[1]

(1. Computer Technology Application Key Laboratory of Yunnan Province, Kunming University of Science and Technology, Kunming 650500, China; 2. Yunnan Observatories, Chinese Academy of Sciences, Kunming 650011, China, Email: wf@cnlab.net; 3. National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100012, China)

Abstract: Mingantu Ultrawide Spectral Radioheliograph (MUSER) generates 100 kilobytes raw observational data in every 3 milliseconds and more than 3. 5 Terabytes data per day. For further data analysis and sharing, it is necessary to convert the raw data stored in self-defined format to standard format usually used in the field of radio astronomy. In previous work, we have analyzed the format of UVFITS and converted the raw data to UVFITS file successfully. However, the efficiency of the format converting system needs to be further improved. This paper presents a parallel UVFITS file assembly system based on cluster parallel environment. Experiments show that the system can reduce the execution time of assembling a UVFITS file to about 1. 2 milliseconds, 2. 5 times faster than that of the data acquisition, which is very promising to meet the data processing requirements in the project. Moreover, the parallel system can be used for reference in other systems. The implementation of this parallel data format converting system can also provide a good reference to similar data processing systems.

Key words: UVFITS; Massive data; MPI; Parallel computing; High-performance computing